CT-Net: Complementary Transfering Network for Garment Transfer with Arbitrary Geometric Changes

Fan Yang, Guosheng Lin*
S-Lab, Nanyang Technological University
School of Computer Science and Engineering, Nanyang Technological University
E-mail: yyyfan3@gmail.com, gslin@ntu.edu.sg

Abstract

Garment transfer shows great potential in realistic applications with the goal of transfering outfits across different people images. However, garment transfer between images with heavy misalignments or severe occlusions still remains as a challenge. In this work, we propose Complementary Transfering Network (CT-Net) to adaptively model different levels of geometric changes and transfer outfits between different people. In specific, CT-Net consists of three modules: i) A complementary warping module first estimates two complementary warpings to transfer the desired clothes in different granularities. ii) A layout prediction module is proposed to predict the target layout, which guides the preservation or generation of the body parts in the synthesized images. iii) A dynamic fusion module adaptively combines the advantages of the complementary warpings to render the garment transfer results. Extensive experiments conducted on DeepFashion dataset demonstrate that our network synthesizes high-quality garment transfer images and significantly outperforms the state-of-art methods both qualitatively and quantitatively. Our source code will be available online.

1. Introduction

Most existing virtual try-on methods are based on simplifying assumptions: (i) Pure clothing images or 3D information are available. (ii) Pose changes are simple without heavy misalignments or severe occlusions. We argue that these simplifying assumptions greatly limit the application scope of these methods in the realistic virtual try-on scenarios. To address this issue, we propose Complementary Transfering Network (CT-Net), a novel image-based garment transfer network that does not rely on pure clothing images or 3D information while capable to adaptively deal with different levels of geometric changes. As shown in Figure 1, given a target person image I^T and a model image I^M, without any restriction to the poses or shapes of I^T and I^M, our CT-Net synthesizes photo-realistic garment transfer results, in which the person in I^T wearing the clothes depicted in I^M with well-preserved details.

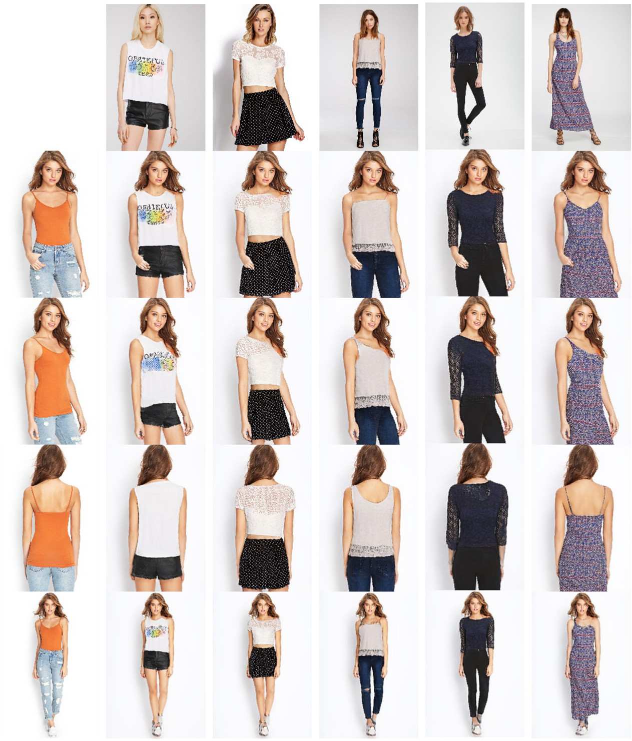

Figure 1. **Garment transfer results generated by CT-Net. First row: model images. First column: target person images.** As shown above, CT-Net naturally transfers clothes across different people with arbitrary poses or shapes and synthesizes photo-realistic images with well-preserved characteristics of the desired clothes and distinct identities of humans. Please refer to *supplementary materials* for more results.

Despite various methods have been proposed to realize

[*] Guosheng Lin is the corresponding author

virtual try-on in different settings [13, 34, 39, 38, 27, 6, 11, 37], there is still a gap between these methods and the unlimited realistic scenarios. Some methods [10, 4, 26] involve 3D information to deal with occlusions, but they are greatly limited by expensive devices and high computational costs. Others [13, 34, 39, 38] may rely on stand-alone clothing images, which are not easy to get timely online. Moreover, most of them attempt to model the geometric changes of the clothes utilizing a Thin Plate Spline (TPS) warping. Because TPS warping is limited by a small number of parameters and only capable to shape simple deformations, their methods fail to deal with complex cases with heavy misalignments or severe occlusions. Garment transfer methods aim to transfer outfits across different people. Although prior arts [27, 11, 37] have achieved considerable progress, none of them address the issue of large geometric changes.

We aim to fulfill this gap by proposing a novel garment transfer network, Complementary Transfering Network (CT-Net), which precisely transfers outfits across different people while tolerating different levels of geometric changes. As shown in Figure 2, CT-Net has three modules:

First, a Complementary Warping Module (CWM) is introduced to warp the desired clothes into the target region. Specially, we simultaneously estimate two complementary warpings with different levels of freedom: (a) Distance fields guided (DF-guided) dense warping. (b) Thin Plate Spline (TPS) warping. DF-guided dense warping has a high degree of freedom and is utilized to warp the desired clothes to be well-aligned with the target pose; while limited by a small number of parameters, TPS warping roughly transfers the desired clothes into the target region with well-preserved textures.

Second, a Layout Prediction Module (LPM) is introduced to predict the target layout, in which the target person wearing the desired clothes. Compared to prior works, which may suffer from the misalignments between inputs [27, 11, 38], our Layout Prediction Module makes more accurate predictions based on the aligned warping results from Complementary Warping Module. Leveraging the predicted target layouts, our network dynamically determines the non-target body areas and the occluded body areas, which guides the adaptive preservation and generation. Benefited from joint training, Layout Prediction Module also adds spatial constraints to the training of complementary warpings, encouraging the warping results to be more coherent with the target person.

Third, a Dynamic Fusion Module (DFM) integrates all the information provided by previous modules to render the garment transfer results. Specifically, our Dynamic Fusion Module adopts an attention mechanism to adaptively combine the advantages of the two complementary warpings and synthesizes photo-realistic garment transfer results with well-preserved characteristics of the clothes.

Extensive experiments conducted on DeepFashion dataset demonstrate the superiority of our method compared to the state-of-art methods. Our main contributions can be summarized as follows:

- We propose a novel image-based garment transfer network, which adaptively combines two complementary warpings to model different levels of geometric changes and synthesizes photo-realistic garment transfer results with well-preserved characteristics of the clothes and distinct human identities.
- A novel Layout Prediction Module makes precise prediction on the target layout, which clearly shapes the synthesized results, guides the adaptive preservation and generation of the body parts and adds spatial constraints to the training of the complementary warpings.
- Evaluated on DeepFashion [21] dataset, CT-Net synthesizes high-quality garment transfer results and outperforms all the state-of-art methods both qualitatively and quantitatively.

2. Related Work

Generative Adversarial Networks. Generative Adversarial Networks (GANs) [9] have been demonstrated very effective in generating fake images, which are indistinguishable from the real ones in the original dataset. Conditional GAN (cGAN) [24] adopts extra information to further control the generation results, which promotes the development of many applications [14, 35, 25, 1, 12]. Specifically, Isola *et al.* [15] proposed an image-to-image translation network to transfer images from one domain to another, which explores relationships across different domains. Similarly, we also employ a cGAN to synthesize photo-realistic garment transfer results conditioned on the desired clothes and the target pose.

Pose-guided Human Image Generation. Ma *et al.* [22] made an early attempt to generate human images conditioned on pose with a two-stage network. Esser *et al.* [7] proposed a conditional U-Net to disentangle the pose and appearance. More recent methods [31, 5, 28, 11] propose to solve this problem utilizing warp-based methods. Zhu *et al.* [41] employs a sequence of pose-attentional transfer blocks to progressively deal with large pose discrepancies. Zhang *et al.* [40] for the first time introduces cross-domain semantic matching, which learns dense correspondence warping between cross-domain inputs. Inspired by [40], we also employ the dense correspondence warping. However, we focus on the exact problem of garment transfer and estimate a distance fields guided dense warping. Benefited from the joint training of all the modules, our distance fields guided dense warping naturally warps the clothing items to be well-aligned with the target pose and preserves

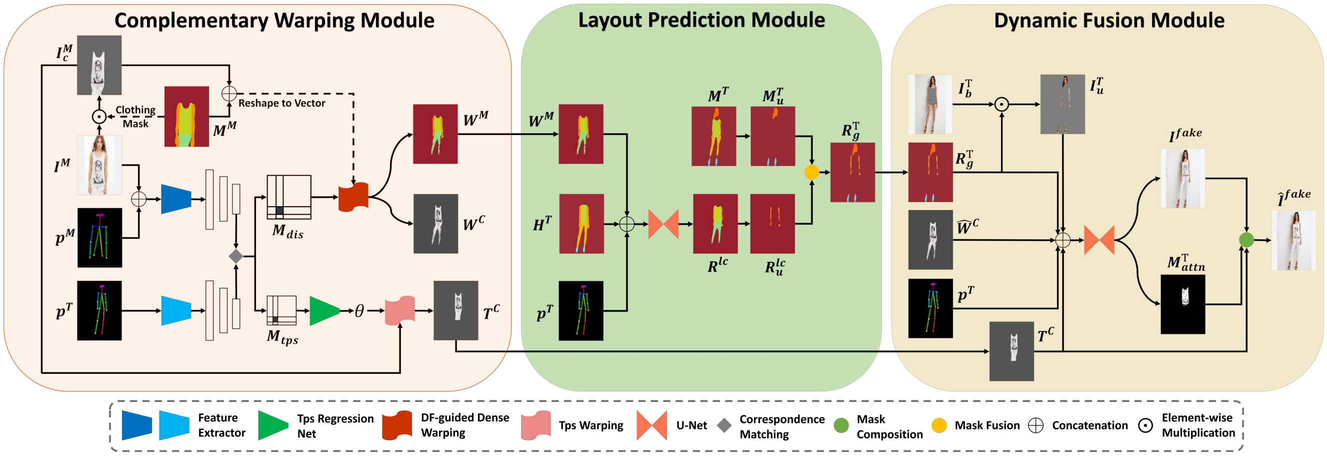

Figure 2. The overall architecture of CT-Net: (i) Complementary Warping Module (CWM) estimates two complementary warpings to warp the desired clothes I_c^M in two different granularities, where $W^{\{\cdot\}}$ denotes the warping result of DF-guided dense warping and $T^{\{\cdot\}}$ denotes the warping result of TPS warping. (ii) Layout Prediction Module (LPM) predicts the target layout R_g^T to guide the layout adaption, where H^T denotes the clothing-agnostic human representations. (iii) Dynamic Fusion Module (DFM) adaptively integrates all the information with an attention mechanism M_{attn}^T to render the photo-realistic garment transfer result $\hat{I}^{fake}$. Note $\widehat{W}^C = W^C \odot R_c^{lc}$, where R_c^{lc} denotes the predicted clothing mask in R^{lc}, which is not shown in the figure for simplicity.

the clothing patterns well.

Virtual Try-on. Many conventional virtual try-on methods rely on 3D information [10, 4, 26]. Along with the advances of deep neural networks, more recent works attempt to synthesize try-on results based on 2D images. Various methods [13, 34, 39, 38, 11] have been proposed to transfer clothes in a stand-alone clothing image onto a target person. However, all of them rely on simplifying assumptions that the clean clothing images are available and geometric changes are simple enough to be modeled by a Thin Plate Spline warping (TPS) with a small number of parameters (*e.g*. 6 for affine and $2 \times 5 \times 5$ for TPS as in [34]). These assumptions greatly hinder the application of these methods in the realistic virtual try-on scenarios. Wu *et al*. [37] proposed to use densepose [2] descriptor to warp the desired clothes onto the target person. But warping estimated by densepose descriptor can be very sparse when there are large occlusions, leading to unconvincing synthesized results. Methods mentioned above only focus on the transfer of upper clothes. SwapNet [27] employs a two-stage network to transfer the entire outfits across people images. To deal with the misalignments of features, they adopt ROI pooling and encode each clothing regions into high-dimensional features. However, the encoded features are inadequate to preserve the local textures, which leads to blurry synthesized results. Different from these methods, we explore a wilder application scope by adaptively combining two complementary warpings to model different levels of deformations and synthesizing high-quality garment transfer results with arbitrary geometric changes.

3. Complementary Transfering Network

Given the image I^M depicting the model wearing desired clothes, the image I^T depicting the target person, assuming clothes, poses and shapes of I^M and I^T can be arbitrary, our goal is to synthesize high-quality garment transfer results with well-preserved characteristics of the clothes and distinct human identities. To achieve our goal, we present Complementary Transfering Network (CT-Net). As shown in Figure 2, CT-Net consists of three modules. First, we introduce a Complementary Warping Module (CWM) to simultaneously estimate two complementary warpings to deal with different levels of geometric changes (Section 3.1). Second, we introduce a Layout Prediction Module (LPM), in which we predict the target layouts to guide the preservation or generation of body parts in the synthesized results (Section 3.2). The third module is a Dynamic Fusion Module (DFM), which adaptively combines the advantages of the complementary warpings to render the garment transfer results (Section 3.3).

3.1. Complementary Warping Module

To synthesize garment transfer results, one of the main challenges is to combine the clothes of the model with the misaligned target pose. A good practice to address this issue is to estimate warpings between the clothes and the target pose [31, 11, 28, 19]. However, to our best knowledge, there is no perfect warping that can shape any geometric changes. Warpings with a high degree of freedom are capable to shape large geometric changes, but they have higher error rates and may fail to preserve complex visual patterns; warping methods with limited numbers of parameters can retain the textures of the desired clothes well, but they can not deal with large geometric changes. Therefore, we propose Complementary Warping Module to simultaneously estimate two complementary warpings with different degrees of freedom to warp the inputs in two granularities, which enables our network to deal with different levels of geometric changes.

As shown in Figure 2, we first employ two separate feature extractors to extract high-level features. Then we match the features to calculate the correspondence matrixs, $\mathcal{M}_{dis}$ and $\mathcal{M}_{tps}$, which are then used to estimate DF-guided dense warping and TPS warping. Given a model image I^M, we estimate the original layout M^M and extract the corresponding clothes I_c^M. Denote the warping results of DF-guided dense warping and TPS warping as $W^{\{\cdot\}}$ and $T^{\{\cdot\}}$. DF-guided dense warping is utilized to transfer I_c^M and M^M in a finer granularity to get W^C and W^M. TPS warping transfers I_c^M to T^C, which is roughly aligned with the target pose.

Pose Representations. We adopt the keypoint distance fields as our pose representations. In detail, we apply the state-of-art pose estimator [3] to estimate keypoint confidence maps of the target person and the model image. Then we convert the sparse joint maps into keypoint distance fields by replacing each zero pixel with its distance to the joint mask. Keypoint distance filed represents each pixel with a unique distance vector, which greatly facilitates the estimation of the correspondence matrixs. In this paper, pose representations of the target person and the model are denoted as p^T and p^M.

Correspondence Matrix. We adopt the keypoint distance fields to estimate dense correspondence matrixs. To be specific, let $\mathcal{F}_A$, $\mathcal{F}_B$ denote the separate feature extractors, we first extract high-level features $m_f \in \mathbb{R}^{H\times W\times C}$ and $t_f \in \mathbb{R}^{H\times W\times C}$ as follows:

$$m_f = \mathcal{F}_A(I^M, p^M), \tag{1}$$

$$t_f = \mathcal{F}_B(p^T), \tag{2}$$

where I^M denotes the model image.

To estimate the correspondence matrixs, we aggregate the features into different scales with different sliding windows, which is illustrated in Figure 3. Specifically, we use sliding window of size 3, with stride 1 and padding size 1 to estimate the correspondence matrix $\mathcal{M}_{dis} \in \mathbb{R}^{HW\times HW}$ for DF-guided dense warping and we apply a sliding window of size 4, with stride 4 and padding size 0 to estimate $\mathcal{M}_{tps} \in \mathbb{R}^{HW/16\times HW/16}$ for TPS warping.

We employ the same correspondence layer as [40] to match the aggregated features, which can be formulated as:

$$\mathcal{M}(i,j) = \frac{(m'_f(i)^T - u_m)(t'_f(j) - u_t)}{\|m'_f(i) - u_m\| \|t'_f(j) - u_t\|}, \tag{3}$$

where m'_f and t'_f represent the aggregated features, u_m and u_t represent the mean vectors.

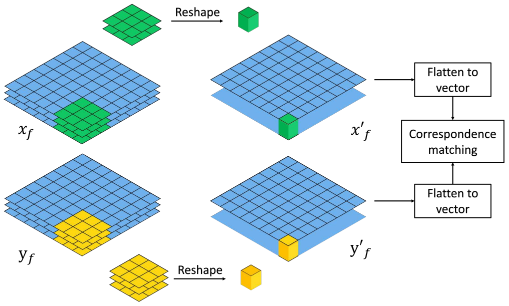

Figure 3. Illustration of the correspondence matching process with sliding window of size 3, stride 1 and padding size 1. Sliding blocks are first extracted from high-level features $\{x_f, y_f\} \in \mathbb{R}^{H\times W\times C}$ and then flattened into a column of the aggregated features $\{x'_f, y'_f\} \in \mathbb{R}^{H\times W\times 9C}$, which are utilized to estimate the correspondence matrix $\mathcal{M} \in \mathbb{R}^{HW\times HW}$.

Distance Fields guided (DF-guided) Dense Warping. According to the dense correspondence matrix $\mathcal{M}_{dis}$, We calculate the weighted average to estimate the distance fields guided (DF-guided) dense warping :

$$\mathcal{W}^X(u) = \sum_v \underset{v}{softmax}(\alpha \mathcal{M}_{dis}(u,v)) \cdot X(v), \tag{4}$$

where α is a hyper-parameter controlling the sharpness of the softmax. We set it as 100 here. DF-guided dense warping learns a dense mapping between two images with a high degree of freedom, which is capable to handle large geometric changes. We utilize it to transfer the clothes of the model image to be well-aligned with the target person, which provides important guides for the generator to reconstruct the local textures of the clothes in the synthesized results.

TPS Warping. Given the model clothes I_c^M, we warp it with the deformation shaped by TPS to be roughly aligned with the target person I^T.

We estimate the TPS warping from $\mathcal{M}_{tps}$. As shown in Figure 2, we first employ a regression net to predict the corresponding control points and then calculate the parameters θ. For training, we adopt the second-order constraint [38] to restrict the TPS warping from generating unnatural deformations or mess textures, which is denoted as $\mathcal{L}_{sc}$. The total loss can be formulated as:

$$\mathcal{L}_{tps} = \lambda_1 \|I_c^T - T^C\|_1 + \lambda_2 \mathcal{L}_{sc}, \tag{5}$$

where I_c^T represents the ground-truth clothes extracted from the model image, λ_1 and λ_2 are the weights for the two loss terms. Both of them are set to 10, respectively, in our experiments.

3.2. Layout Prediction Module

We propose Layout Prediction Module to predict target layouts, in which the target person wearing the desired clothes depicted in the model image. Prior works [27, 5, 11] mostly generate the target layouts conditioned on the original layouts of the model and the target pose. However, suffered from the limited receptive fields of CNNs, these methods fail to understand the correlation between the original layout and the pose representations while facing with large geometric changes. Compared to these methods, we explore a warping-based strategy to eliminate the misalignments and facilitates the prediction.

As shown in Figure 2, we first align the original layout with the target pose leveraging the DF-guided dense warping and then feed the warped layout W^M with the clothing-agnostic representations H^T into the U-net [29] to predict the target limb and clothing mask R^{lc}. Denote the head and shoes mask of the target person as M_u^T, the limb mask of the predicted layout as R_u^{lc}, we merge M_u^T with R_u^{lc} to form the complete target layouts R_g^T, which provides important guides for the generator to adaptively determine the preservation or generation of the body parts in the synthesized results (Section 3.3).

We get the original layouts utilizing the state-of-art human parsing network [20] and adopt the segmentation of densepose descriptor [2] as our extra clothing-agnostic representations H^T. Since head and shoes are not in the transfer list of our model, we remove these areas in H^T and the model's layout M^M. For training this module, we adopt the pixel-level cross-entropy loss, denoted as $\mathcal{L}_{layout}$. Benefited from joint training, Layout Prediction Module adds extra spatial constraints to the training of Complementary Warping Module, encouraging the warping results to be more coherent with the target person.

3.3. Dynamic Fusion Module

Dynamic Fusion Module is proposed to adaptively combine the advantages of the two complementary warpings to render the garment transfer results with well-preserved characteristics of the clothes and distinct identities of humans. As shown in Figure 2, we first utilize the merged target layout R_g^T to extract the non-target body parts from the non-clothing model image I_b^T. Leveraging the non-target body parts and the target layout, the generator learns to preserve the details in the non-target body parts and inpaint occluded body parts according to the target layouts, leading to well-preserved human identities and clear body boundaries in the synthesized results. We then adopt a cGAN to integrate the non-target body parts I_u^T, the target layout R_g^T, complementary warping results $\widehat{W}^C$, T^C and target pose representation p^T to render the initial generation result I^{fake}. Note $\widehat{W}^C = W^C \odot R_c^{lc}$, where R_c^{lc} denotes the predicted clothing mask in R^{lc}, which is not shown in the Figure 2 for simplicity.

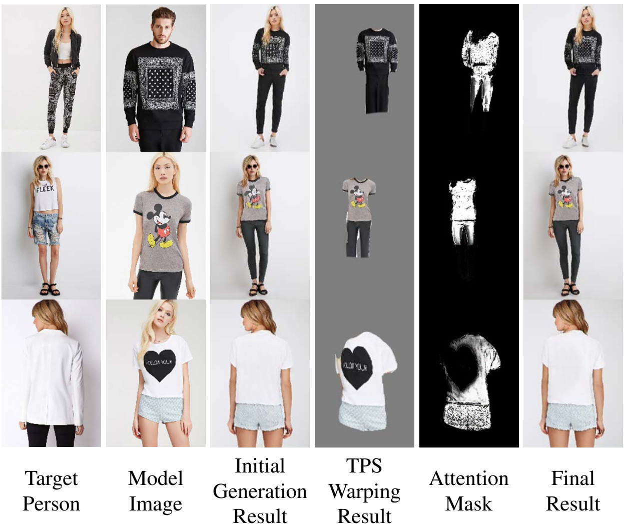

Figure 4. Examples of our attention mechanism. From left to right: target person I^T, model image I^M, initial generation result I^{fake}, TPS warping result T^C, attention mask M_{attn}^T, final result $\hat{I}^{fake}$.

An attention mechanism is employed to combine the advantages of the two complementary warpings, where an attention mask M_{attn}^T is estimated to compose the initial generation result I^{fake} with the warping result from TPS as our final garment transfer result $\hat{I}^{fake}$:

$$\hat{I}^{fake} = T^C \odot M_{attn}^T + I^{fake} \odot (1 - M_{attn}^T). \quad (6)$$

As shown in Figure 4, our attention mechanism adaptively selects different regions from the initial generation result I^{fake} and the warping result of TPS according to different levels of geometric changes. For example, when the geometric change is simple and can be shaped by the TPS warping, as the first two rows in Figure 4, the attention mechanism selects more regions on the warping result from TPS to refine the initial generation result; while as the third row, when there are heavy misalignments or severe occlusions, the attention mechanism tends to retain the initial generation result and ignore the large logos or unreasonable textures in the warping result of TPS. In this way, we adaptively combines the two complementary warpings to deal with different levels of geometric changes and expand the application scope of our model to wilder scenarios.

3.4. Loss Functions

To encourage the training of different modules benefit each other, we train our model in a joint style. We combine several different losses to produce high-quality garment transfer results:

Perceptual Loss. Based on the difference between high-level features, perceptual loss has been proved efficient in the image generation tasks [17]. To pose perceptual constraints on the synthesized results, we adopt a pre-trained VGG network [32] to extract multi-level features ϕ_j and compute the perceptual loss as:

$$\mathcal{L}_{perceptual} = \sum_{j=1}^{N} \lambda_j \|\phi_j(\hat{x}_t) - \phi_j(x)\|_2. \tag{7}$$

Style Loss. We further apply the style loss [8] to penalize the statistic error between high-level features, which can be formulated as:

$$\mathcal{L}_{style} = \sum_{j=1}^{N} \|G_j^{\phi}(\hat{x}_t) - G_j^{\phi}(x_t)\|_2, \tag{8}$$

where G_j^{ϕ} denotes the Gram matrix estimated from ϕ_j.

Contextual Loss. To encourage our network to preserve more details from the desired clothes I_c^M, we employ the contextual loss proposed in [23], which can be formulated as:

$$\mathcal{L}_{contextual} = \sum_{l=1} \lambda_l \left[-\log \left(\frac{1}{n_l} \sum_i \max_j A^l(\phi_i^l(\hat{x}_B), \phi_j^l(y_B)) \right) \right], \tag{9}$$

where A^l denotes the pairwise affinities between features.

Adversarial Loss. To force the generator to learn the real distributions of the dataset and generate realistic human images, we deploy a discriminator to discriminate the generated fake images from the real samples in the dataset. The loss can be formulated as:

$$\mathcal{L}_{adv} = \mathbb{E}_{x,y}[\log(\mathcal{D}(x,y))] + \mathbb{E}_x[\log(1 - \mathcal{D}(x, \mathcal{G}(x)))], \tag{10}$$

where x represents the inputs and y is the ground-truth.

Objective Function. Besides the losses above, we apply a L1 regularization $\mathcal{L}_{reg} = \|1 - M\|_1$ on M_{attn}^T to prevent the network from overfitting to the initial synthesized result I^{fake}. We also take a L1 loss to stabilize our training process, which can be defined as $\mathcal{L}_{l1} = \|\hat{x} - x\|_1$. Our objective function is a weighted sum of above terms:

$$\mathcal{L}_{total} = \alpha_1 \mathcal{L}_{l1} + \alpha_2 \mathcal{L}_{tps} + \alpha_3 \mathcal{L}_{layout} + \alpha_4 \mathcal{L}_{perceptual} + \alpha_5 \mathcal{L}_{style} + \alpha_6 \mathcal{L}_{contextual} + \alpha_7 \mathcal{L}_{adv} + \alpha_8 \mathcal{L}_{reg}, \tag{11}$$

where α_i,$(i = 1, \ldots, 8)$ are hyper-parameters controlling the weights of each loss.

4. Experiments

4.1. Dataset

We evaluate our model on the In-shop Clothes Retrieval Benchmark of DeepFashion dataset [21], which contains 52,712 fashion images of resolution 256 × 256. For training, we select 37,836 pairs of images depicting the same person wearing the same outfit with different poses. At test stage, we select 4,932 pairs of images which are not overlapped with the training set. As the realistic virtual try-on scenarios, each testing pair contains two different people with different clothes and poses.

4.2. Implementation Details

We adopt Adam [18] with $\beta_1 = 0.5, \beta_2 = 0.999$ as the optimizer in our all experiments. Our model is trained in stages. Complementary Warping Module is firsted trained for 20 epoches to estimate reasonable warpings. Then our model is jointly trained in an end-to-end manner for another 80 epoches. Learning rate is fixed at 0.0002 for the first 40 epoches and then decays to zero linearly in the remaining steps. InstanceNorm2d Normalization [33] is applied to all layers to stabilize the training. The detailed network structures can be found in the *supplementary materials*. To balance the scales of losses in Eqn. 11, we set $\alpha_{1,2,3,7,8} = 10$ and $\alpha_{4,5,6} = 1$.

4.3. Baselines

ACGPN. ACGPN is a state-of-art virtual try-on network proposed by Yang *et al.* [38], which aims to transfer a stand-alone clothing image onto a reference person. In comparison to previous methods [13, 34], ACGPN first predicts the target clothing segmentation progressively in two stages, then estimates a TPS warping utilizing STN [16]. ACGPN shows state-of-art performance on VITON [13] dataset with natural deformed clothes and well-preserved non-target body parts.

CoCosNet. CoCosNet stands for the cross-domain correspondence network proposed by Zhang *et al.* [40], aiming to synthesize realistic images according to the examplar images. Different from other methods, CoCosNet establishes a cross-domain correspondence matching to align the examplar image with the target image and then synthesize photo-realistic results, which achieves state-of-art performance in pose-guided human generation task. However, CoCosNet lacks of the ability to generate garment transfer results. To adapt it into our task, we apply the cross-domain correspondence warping to warp the layouts and replace the inputs for the Translation Network in CoCosNet to be the same as ours in the Dynamic Fusion Module.

To keep the fairness of our experiments, We retrain all aforementioned methods on DeepFashion [21] dataset with the same training set as ours.

4.4. Qualitative Results

Figure 5 shows the qualitative comparisons of ACGPN, CoCosNet and our model, which indicates that our model synthesizes more convincing results with well-preserved

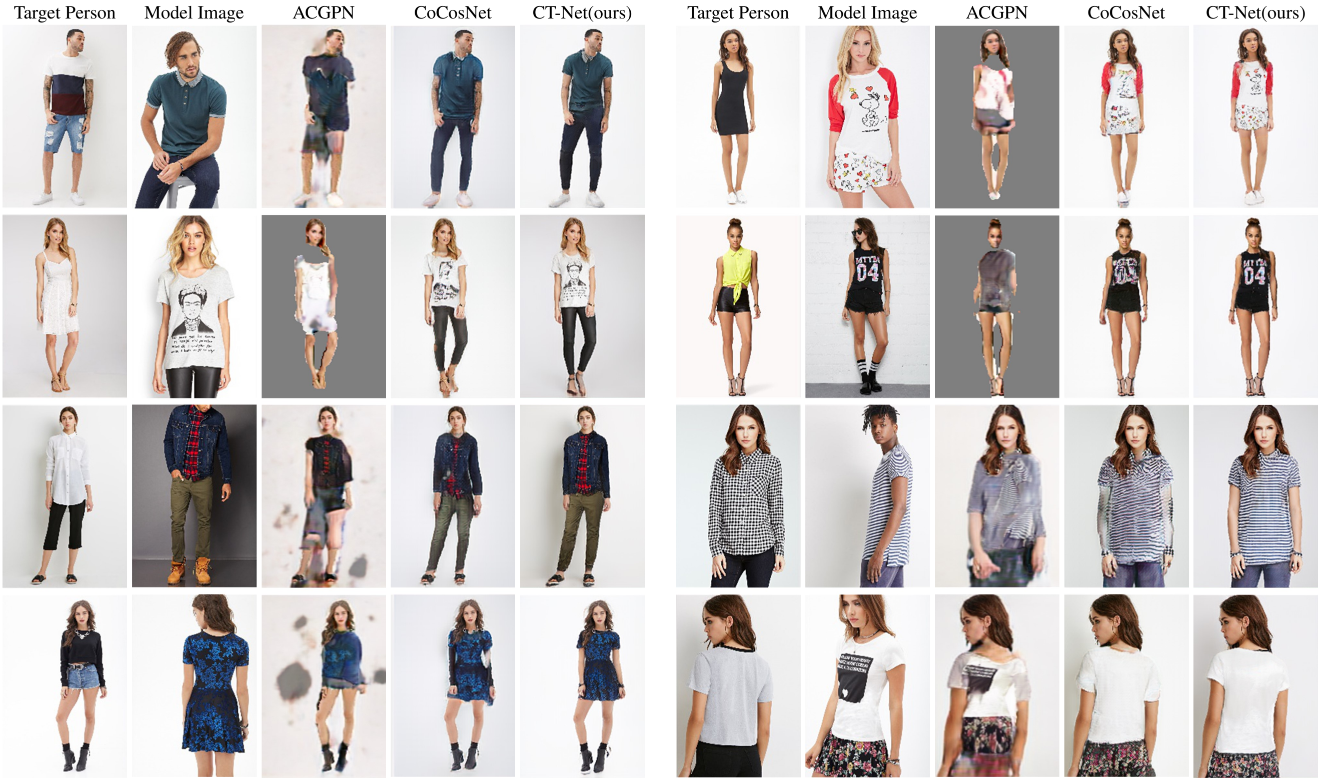

Figure 5. Qualitative comparisons of our method with our baselines.

characteristics of clothes and distinct identities of humans. Since ACGPN overlooks the misalignments between the inputing segmentations, it fails to generate the correct target clothing segmentation and estimate reasonable TPS warping, leading to unsatisfying results with messy textures, incorrect body parts and abundant artifacts. Our CoCosNet [40] baseline eliminates the misalignments by employing the cross-domain correspondence warping. However, cross-domain correspondence warping has a high degree of freedom and fails to preserve the complex clothing patterns, result in visual artifacts such as cluttered textures and blurry boundaries. Benefited from the Layout Prediction Module, our CT-Net adaptively preserves the non-target body parts and generates the occluded parts, leading to realistic garment transfer results with distinct human identities and clear body boundaries. As shown in the first row (right) and second row of the Figure 5, the proposed attention mechanism in the Dynamic Fusion Module adaptively combines the advantages of the two complementary warpings and preserves the logos on the desired clothes clearly. In the last row (right), since ACGPN is unaware of the change of the human pose, it wrongly preserves the logo in the back view of the person, while our attention mechanism adaptively drops the logo and synthesizes reasonable back-view image.

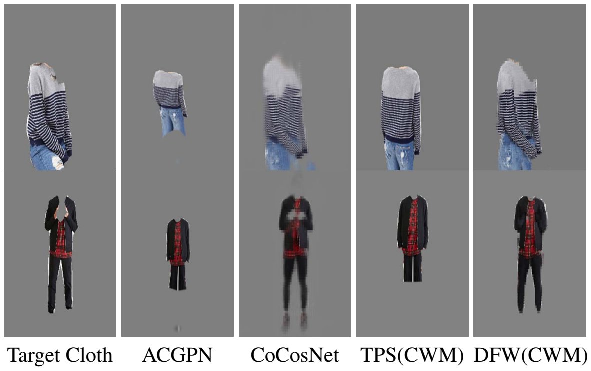

Figure 6. Visual comparisons of warping results. DFW(CWM) represents the warping results from DF-guided dense warping estimated in the Complementary Warping Module.

In Figure 6, we visualize warping results from different methods to make further comparisons. Benefited from the joint training of all modules, CT-Net shows superior performance in the estimation of both the DF-guided dense warping and the TPS warping. Please refer to *supplementary materials* for more qualitative results.

4.5. Quantitative Results

We adopt Structural Similarity (SSIM) [36] to measure the similarity between the generated images and the real

Methods	Warp-SSIM	Mask-SSIM	H-SSIM	IS
ACGPN [38]	0.744	0.757	0.813	3.366
CoCosNet [40]	0.815	0.835	0.851	3.472
w/o PR	0.857	0.913	0.919	3.495
w/o LPM	0.836	0.917	0.923	3.479
w/o TPS	0.860	0.919	**0.931**	**3.515**
CT-Net (ours)	**0.865**	**0.923**	0.930	3.511

Table 1. Quantitative comparisons of our method with other methods.

ones. To compute SSIM, we take the same person wearing the same clothes but in different poses as test set. Specifically, inspired by [31, 11], we compute SSIM in three different scales: (i) To isolate the influence from the background, we compute SSIM for human pixels (H-SSIM). (ii) To evaluate the accuracy of reconstructed clothes, we compute SSIM for clothing area in the synthesized results (Mask-SSIM). (iii) To compare the warping accuracy of different methods, we compute SSIM for warped clothes (Warp-SSIM). For our model, we use the warping results from DF-guided dense warping to calculate Warp-SSIM. Besides, we adopt Inception Score (IS) [30] to evaluate the quality of our synthesized images.

Table 1 reports the quantitative results of our method and baselines. Higher scores are better. As summarized in table 1, our method outperforms all baselines by a significant margin. Specially, our model greatly improves the warping accuracy with 0.050 higher Warp-SSIM scores compared to CoCosNet. Moreover, we also achieve higher scores in terms of H-SSIM, Mask-SSIM and IS, which indicates that our method synthesizes more realistic images with well-preserved details.

4.6. Ablation Study

We conduct ablation experiments to explore the effectiveness of the main components in our model. In particular, *w/o* PR denotes removing the pose representation p^M inputing to the feature extractor in Complementary Warping Module. *w/o* LPM denotes removing Layout Prediction Module. *w/o* TPS denotes removing the estimation of TPS warping.

Table 1 reports all the results of our ablation experiments. In specific, our full model outperforms all ablation methods by a margin in Warp-SSIM, which indicates that our designs in the network significantly facilitate the estimation of DF-guided dense warping. Combining the advantages of the two complementary warpings, our full model also shows the best performance in reconstructing the clothes and achieves the highest Mask-SSIM scores. Our full model and *w/o* TPS have similar scores in all metrics, since SSIM only roughly measures the local similarity and IS only captures the realism of the images.

To further demonstrate the superiority of our full model, we visualize some examples to make qualitative comparisons in Figure 7. Since *w/o* PR and *w/o* LPM can not estimate the warping precisely, artifacts such as incorrect clothing shape (first row) and blurry boundaries (third row) can be observed. Although *w/o* TPS achieves the best scores in terms of H-SSIM and IS, visual results show that our full model synthesizes more photo-realistic images with better-preserved clothing patterns and distinct body parts.

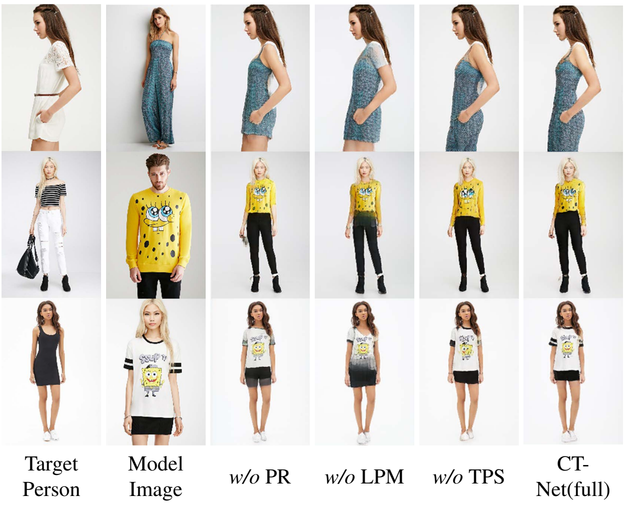

Figure 7. Visual comparisons with ablation methods.

5. Conclusion

We propose Complementary Transfering Net (CT-Net) for garment transfer with arbitrary geometric changes. In particular, our model adaptively combines two complementary warpings to model different levels of geometric changes and synthesizes photo-realistic garment transfer results with well-preserved characteristics of the clothes and distinct human identities. We introduce three novel modules: i) Complementary Warping Module. ii) Layout Prediction Module. iii) Dynamic Fusion Module. Experiment results demonstrate that our model significantly outperforms state-of-art methods both qualitatively and quantitatively.

6. Acknowledgement

This research was conducted in collaboration with SenseTime. This work is supported by A*STAR through the Industry Alignment Fund - Industry Collaboration Projects Grant. This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-RP-2018-003), and the MOE Tier-1 research grants: RG28/18 (S) and RG22/19 (S).

References

[1] Badour AlBahar and Jia-Bin Huang. Guided image-to-image translation with bi-directional feature transformation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9016–9025, 2019. 2

[2] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 3, 5

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 4

[4] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Synthesizing training images for boosting human 3d pose estimation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 479–488. IEEE, 2016. 2, 3

[5] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. In *Advances in neural information processing systems*, pages 474–484, 2018. 2, 5

[6] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. Fw-gan: Flow-navigated warping gan for video virtual try-on. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1161–1170, 2019. 2

[7] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018. 2

[8] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Advances in neural information processing systems*, pages 262–270, 2015. 6

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[10] Peng Guan, Loretta Reiss, David A Hirshberg, Alexander Weiss, and Michael J Black. Drape: Dressing any person. *ACM Transactions on Graphics (TOG)*, 31(4):1–10, 2012. 2, 3

[11] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10471–10480, 2019. 2, 3, 5, 8

[12] Xintong Han, Zuxuan Wu, Weilin Huang, Matthew R Scott, and Larry S Davis. Finet: Compatible and diverse fashion image inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4481–4491, 2019. 2

[13] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018. 2, 3, 6

[14] Wei-Lin Hsiao, Isay Katsman, Chao-Yuan Wu, Devi Parikh, and Kristen Grauman. Fashion++: Minimal edits for outfit improvement. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5047–5056, 2019. 2

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 6

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 6

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[19] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3693–3702, 2019. 3

[20] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):871–885, 2018. 5

[21] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 2, 6

[22] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in neural information processing systems*, pages 406–416, 2017. 2

[23] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 768–783, 2018. 6

[24] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[25] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 2

[26] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (TOG)*, 36(4):1–15, 2017. 2, 3

[27] Amit Raj, Patsorn Sangkloy, Huiwen Chang, James Hays, Duygu Ceylan, and Jingwan Lu. Swapnet: Image based garment transfer. In *European Conference on Computer Vision*, pages 679–695. Springer, 2018. 2, 3, 5

[28] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7690–7699, 2020. 2, 3

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5

[30] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016. 8

[31] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3408–3416, 2018. 2, 3, 8

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[33] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 6

[34] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 589–604, 2018. 2, 3, 6

[35] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 2

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[37] Zhonghua Wu, Guosheng Lin, Qingyi Tao, and Jianfei Cai. M2e-try on net: Fashion from model to everyone. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 293–301, 2019. 2, 3

[38] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7850–7859, 2020. 2, 3, 4, 6, 8

[39] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10511–10520, 2019. 2, 3

[40] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020. 2, 4, 6, 7, 8

[41] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2347–2356, 2019. 2